

LIMITS OF PERFORMANCE OF REAL-TIME DNA MICROARRAYS

Haris Vikalo¹ and Babak Hassibi²

¹Department of Electrical and Computer Engineering, The University of Texas, Austin, TX

²Department of Electrical Engineering, California Institute of Technology, Pasadena, CA

ABSTRACT

DNA microarrays rely on chemical attraction between the nucleic acid sequences of interest (mRNA and DNA sequences, referred to as *targets*) and their molecular complements which serve as biological sensing elements (*probes*). The attraction between the complementary sequences leads to binding, in which probes capture target molecules. Molecular binding is a stochastic process and hence the number of captured analytes at any time is a random variable. Today, majority of DNA microarrays acquire only a single measurement of the binding process, essentially taking one sample from the steady-state distribution of the binding process. *Real-time* DNA microarrays provide much more: they can take multiple temporal measurements which not only allow more precise characterization of the steady-state but also enable faster detection based on the early kinetics of the binding process. In this paper, we derive the Cramer-Rao lower bound on the mean-square error of estimating the target amounts in real-time DNA microarrays, and compare it to that of conventional microarrays. The results suggest that a few temporal samples collected in the early phase of the binding process are often sufficient to enable significant performance improvement of the real-time microarrays over the conventional ones.

Index Terms: DNA microarrays, Cramer-Rao lower bound

1. INTRODUCTION

DNA microarrays detect presence and quantify amounts of various nucleic acid sequences (DNA, mRNA, microRNA) [1, 2, 3, 4]. They measure interaction between molecules of interest (target analytes) and biological sensing elements, and generate signal proportional to the amount of target molecules. Detection in DNA microarrays relies on the chemical attraction between target molecules and their Watson-Crick complements, which serve as biological sensing elements (probes). The attraction between the molecules leads to binding, in which probes capture target analytes. A transducer then converts the number of complex molecular structures that are formed due to the binding into a signal.

This work was supported by the National Science Foundation under grant CCF-0845730.

DNA microarrays are capable of screening tens or even hundreds of thousands of different gene sequences at the same time, revealing critical information about the functionality of cells, effects of drugs on organisms, etc. They are time and cost efficient, and may enable exciting new applications in drug discovery, medicine, defense systems, and environmental monitoring. To fully realize these promises, however, the performance of microarrays needs to be significantly improved. The sensitivity, dynamic range, and resolution of DNA microarrays is limited by interference, noise, probe saturation, and other sources of errors in the analyte detection procedure [5, 6, 7]. Several of these limitations stem from the fact that the molecular binding is a stochastic process, which majority of the conventional DNA microarrays attempt to characterize based on a single measurement of its equilibrium, i.e., by taking one sample from the steady-state distribution of the binding process. Kinetics of the binding process is illustrated in Fig. 1.

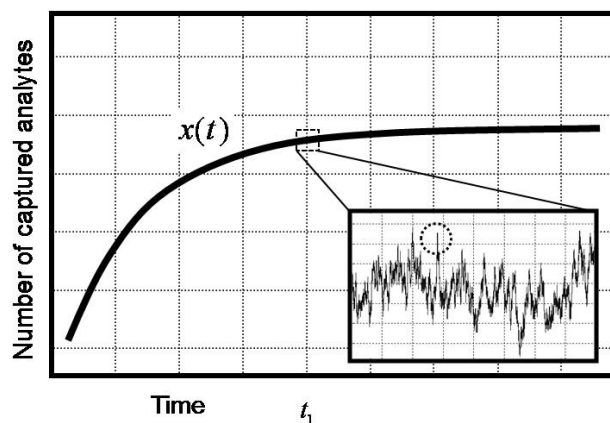


Figure 1: *Kinetics of the molecular binding process. The number of captured target analytes at any point in time is a random variable.*

On the other hand, real-time biosensors are capable of taking multiple temporal samples of a binding process [8, 9, 10]. However, analyte estimation therein is typically performed using only the data collected in the equilibrium, and rarely relies on the kinetics [11]. In [12, 13], analyte

targets in real-time DNA microarrays are estimated using the temporally sampled kinetics of the binding process. It was demonstrated there empirically that the real-time data acquisition may significantly improve the performance of DNA microarrays.

In this paper, we analytically study the achievable mean-square error (MSE) performance of the target analyte estimation in real-time DNA microarrays. Specifically, we compute the Cramer-Rao lower bound (CRLB) on the achievable mean-square estimation error therein, and compare it with the CRLB on the performance of the conventional microarray systems. The paper is organized as follows. Section 2 describes a mathematical model of the real-time microarrays, based on which the Cramer-Rao lower bound is computed in Section 3. To compare the CRLB of a real-time microarray with that of a conventional microarray, an illustrative example is analyzed in Section 4. The paper is concluded in Section 5.

2. MATHEMATICAL MODEL

Consider a DNA microarray with N different types of probes on its surface. Each probe is designed to capture one of the targets possibly present in a biological sample that is to be tested. Let M denote the actual number of different target types that are present in the sample, $M \leq N$. The real-time DNA microarrays acquire temporal samples of the binding process, i.e., they provide a time-series of the number of captured target molecules collected at discrete points in time. Note that in addition to hybridization to its matching probe, a target molecule may also engage in non-specific cross-hybridization with probes whose nucleotide sequences are only partial matches with the target. We assume that, in general, each target molecule of type i may hybridize to its corresponding specific probe as well as cross-hybridize to $C_i \leq N - 1$ non-specific ones.

Both hybridization and cross-hybridization are treated as random events. The probabilities of the specific and non-specific binding of the i^{th} target to the i^{th} and j^{th} probe are denoted by p_{ii} and p_{ij} , respectively. On the other hand, if all we have in the system is binding (i.e., hybridization and cross-hybridization) then, if enough probes are present, eventually all target molecules would bind to the probes. However, this is not the case since both hybridization and cross-hybridization are reversible processes: once a target molecule is bound to a probe there is a nonzero probability that it will be released. We denote the release probability from the hybridized and cross-hybridized state by p_{ii}^r and p_{ij}^r , respectively. Possible states of a target molecule are illustrated in Fig. 2, where we assume one hybridization and one cross-hybridization (states 'H' and 'C', respectively), and where a free target target molecule is in state 'F'.

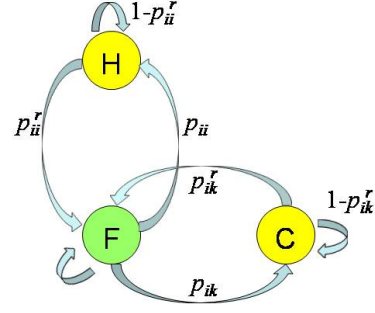


Figure 2: An illustration of a simple Markov chain model describing the possible states of a target molecule. One hybridization (state 'H') and, for simplicity, one cross-hybridization (state 'C') are assumed. A free target molecule is in state 'F'.

We introduce the following notation,

- $n_i(t) \triangleq$ the number of free targets of type i at time t ,
- $n_{b,ij}(t) \triangleq$ the number of targets of type i bound to the probe j at time t ,
- $n_{b,j}(t) \triangleq$ the total number of targets of all types captured by the probe j .

Note that $n_i(0)$ is the total number of target molecules of type i in the biological sample being tested. Moreover, it holds that

$$n_i(t) = n_i(0) - \sum_j n_{b,ij}(t), \text{ and } n_{b,j}(t) = \sum_i n_{b,ij}(t).$$

Assume the realistic scenario where the number of target molecules is much smaller than the number of probe molecules. In other words, assume that there is no saturation of the probes. Then the change in the number of target molecules of type i bound to the probe molecules of type j in the time interval $(t, t + \Delta)$ is given by

$$n_{b,ij}(t+\Delta) - n_{b,ij}(t) = (n_i(0) - \sum_{j'} n_{b,ij'}(t))p_{ij} - n_{b,ij}p_{ij}^r.$$

Now, the binding probability p_{ij} is affected by the capturing process – the fewer available (i.e., unbound) probes, the less likely that a free target will be captured. Let π_{ij} denote the probability of a target molecule of type i being captured by the probe j when the number of probe molecules is unlimited. Then the probability p_{ij} in $(t, t + \Delta)$ is given by

$$p_{ij} = \frac{n_p - \sum_{i'} n_{b,i'j}(t)}{n_p} \pi_{ij},$$

where n_p denotes the number of probe molecules. Therefore, we can write $n_{b,ij}(t + \Delta) - n_{b,ij}(t) =$

$$(n_i(0) - \sum_{j'} n_{b,ij'}(t))(1 - \frac{\sum_{i'} n_{b,i'j}(t)}{n_p})\pi_{ij} - n_{b,ij}p_{ij}^r.$$

By dividing both sides with Δ and letting $\Delta \rightarrow 0$, we obtain

$$\begin{aligned} \frac{dn_{b,ij}}{dt} &= (n_i(0) - \sum_{j'} n_{b,ij'}(t))(1 - \frac{\sum_{i'} n_{b,i'j}(t)}{n_p})k_{ij} \\ &- n_{b,ij}(t)k_{ij}^r + w_{ij}(t), \end{aligned} \quad (1)$$

where k_{ij} and k_{ij}^r denote the forward and backward rate reaction of the binding and disassociation process, respectively, and where we introduced w_{ij} to model the uncertainty of the sensing process. The random variables w_{ij} are assumed to be zero-mean Gaussian with variance proportional to $n_{b,ij}$, i.e., the uncertainty has shot-noise characteristics [14]. We focus on the early part of the reaction where $\sum_{j'} n_{b,ij'} \ll n_i(0) \ll n_p$. Hence we may ignore the quadratic terms in (1), and after rearranging obtain

$$\begin{aligned} \frac{dn_{b,ij}}{dt} &= - \left(\sum_{j'} n_{b,ij'}(t) + \frac{n_i(0)}{n_p} \sum_{i'} n_{b,i'j}(t) \right) k_{ij} \\ &- n_{b,ij}(t)k_{ij}^r + n_i(0)k_{ij} + w_{ij}(t). \end{aligned} \quad (2)$$

For convenience, denote $\mathbf{x}_i = [n_{b,i1} \ n_{b,i2} \dots n_{b,iN}]^T$, $1 \leq i \leq M$, and $\mathbf{x} = [\mathbf{x}_1 \ \mathbf{x}_2 \dots \mathbf{x}_M]^T$. So, \mathbf{x} is an MN -dimensional vector comprising amounts of M target types bound to each of the N probes. Let $\mathbf{k}_i = [k_{i1} \ k_{i2} \dots k_{iN}]^T$, and $\mathbf{k}_i^r = [k_{i1}^r \ k_{i2}^r \dots k_{iN}^r]^T$. Moreover, let

$$\begin{aligned} D_{\mathbf{k}} &= \text{diag}(\mathbf{k}_1, \mathbf{k}_2, \dots, \mathbf{k}_M), \text{ and} \\ D_{\mathbf{k}^r} &= \text{diag}(\mathbf{k}_1^r, \mathbf{k}_2^r, \dots, \mathbf{k}_M^r), \end{aligned}$$

i.e., $D_{\mathbf{k}}$ and $D_{\mathbf{k}^r}$ are diagonal matrices having k_{ij} and k_{ij}^r as the diagonal entries, respectively. Then we can write (2) as

$$\frac{d\mathbf{x}}{dt} = \mathbf{b} - A\mathbf{x} + \mathbf{w}, \quad (3)$$

where the MN -dimensional vectors \mathbf{b} and \mathbf{w} are defined as

$$\mathbf{b} = \begin{bmatrix} n_1(0)\mathbf{k}_1 \\ n_2(0)\mathbf{k}_2 \\ \vdots \\ n_M(0)\mathbf{k}_M \end{bmatrix}, \quad \mathbf{w} = \begin{bmatrix} w_{11} \\ w_{12} \\ \vdots \\ w_{MN} \end{bmatrix},$$

and where

$$A = D_{\mathbf{k}^r} + D_{\mathbf{k}}[(\mathbf{1}\mathbf{1}^T) \otimes I] + \frac{1}{n_p} D_{\mathbf{b}}[I \otimes (\mathbf{1}\mathbf{1}^T)]. \quad (4)$$

The $MN \times MN$ matrix $D_{\mathbf{b}}$ has \mathbf{b} on the diagonal, and zeros everywhere else. The solution to (3) is of the form

$$\mathbf{x}(t) = (I - e^{-At})A^{-1}\mathbf{b} + \underbrace{\int_0^t e^{-A(t-\tau)}\mathbf{w}(\tau)d\tau}_{\mathbf{u}(t)}. \quad (5)$$

Note that the number of target molecules captured by each of the probes can be expressed as

$$\begin{bmatrix} n_{b,1}(t) \\ \vdots \\ n_{b,N}(t) \end{bmatrix} = \begin{bmatrix} \sum_i n_{b,i1}(t) \\ \vdots \\ \sum_i n_{b,iN}(t) \end{bmatrix} = H\mathbf{x}(t),$$

where the dimension of $H = [I_N \ I_N \dots I_N]$ is $N \times MN$. The measured process is thus given by

$$\mathbf{y}(t) = H\mathbf{x}(t) + \mathbf{v}(t).$$

The real-time microarrays collect temporal samples of the process $\mathbf{y}(t)$, i.e., they acquire

$$\mathbf{y}(i\Delta) = H\mathbf{x}(i\Delta) + \mathbf{v}(i\Delta), \quad 1 \leq i \leq L, \quad (6)$$

where $E\{\mathbf{v}(i\Delta)\mathbf{v}(j\Delta)^T\} = \sigma_{\mathbf{v}}^2 I_N \delta_{ij}$. Let

$$\mathbf{y} = [y(\Delta)^T \ y(2\Delta)^T \ \dots \ y(L\Delta)^T]^T.$$

We are interested in finding how precisely the amounts of targets

$$\theta = [n_1(0) \ n_2(0) \ \dots \ n_M(0)]^T$$

can be estimated from the observed \mathbf{y} .

Note that the variance of $\mathbf{u}(t)$ in (5) is $\mathcal{R}_{\mathbf{u}}(t, t') =$

$$\begin{aligned} &= E\{\mathbf{u}(t)\mathbf{u}^T(t')\} \\ &= E\int_0^t \int_0^{t'} e^{-A(t-\tau)} w(\tau) w^T(\tau') e^{-A^T(t'-\tau')} d\tau d\tau' \\ &= \int_0^t \int_0^{t'} e^{-A(t-\tau)} E\{w(\tau)w^T(\tau')\} e^{-A^T(t'-\tau')} d\tau d\tau' \\ &= \int_0^t \int_0^{t'} e^{-A(t-\tau)} \mathcal{R}_{\mathbf{w}}(\tau) \delta(\tau - \tau') e^{-A^T(t'-\tau')} d\tau d\tau' \\ &= \int_0^t e^{-A(t-\tau)} \mathcal{R}_{\mathbf{w}}(\tau) e^{-A^T(t'-\tau)} d\tau, \end{aligned}$$

where $\mathcal{R}_{\mathbf{w}}(\tau) = D_{(I - e^{-A\tau})A^{-1}\mathbf{b}}$, and we assume without a loss of generality $t' > t$. The covariance of the observations \mathbf{y} is given by an $LN \times LN$ matrix R comprising $L^2 \ N \times N$ -dimensional block entries $R_{l_1 l_2}$, $1 \leq l_1, l_2 \leq L$ computed as

$$R_{l_1 l_2} = H R_{\mathbf{u}, l_1 l_2} H^T + \sigma_{\mathbf{v}}^2 I_N \delta_{l_1 l_2},$$

where, for $l_2 \geq l_1$,

$$\begin{aligned} R_{\mathbf{u}, l_1 l_2} &= \mathcal{R}_{\mathbf{u}}(l_1 \Delta, l_2 \Delta) \\ &= \int_0^{l_1 \Delta} e^{-A(l_1 \Delta - \tau)} \mathcal{R}_{\mathbf{w}}(\tau) e^{-A^T(l_2 \Delta - \tau)} d\tau. \end{aligned} \quad (7)$$

Note that in the early phase of a binding reaction we can approximate $e^{-At} \approx I - At$. Then

$$R_{\mathbf{u}, l_1 l_2} = \int_0^{l_1 \Delta} [I - A(l_1 \Delta - \tau)] \tau D_{\mathbf{b}} [I - A^T(l_2 \Delta - \tau)] d\tau.$$

Straightforward integration gives

$$\begin{aligned}
R_{\mathbf{u}, l_1 l_2} &= \frac{1}{2} D_{\mathbf{b}} (l_1 \Delta)^2 - D_{\mathbf{b}} A^T \left[\frac{1}{2} \Delta^2 l_1 l_2 - \frac{1}{3} (l_1 \Delta)^3 \right] \\
&- AD_{\mathbf{b}} \left[\frac{1}{2} (l_1 \Delta)^2 - \frac{1}{3} (l_1 \Delta)^3 \right] \\
&+ AD_{\mathbf{b}} A^T \left[\frac{1}{6} \Delta^4 l_1^3 l_2 - \frac{1}{12} \Delta^4 l_1^4 \right] \quad (8)
\end{aligned}$$

3. CRAMER-RAO LOWER BOUND

Cramer-Rao lower bound gives the best possible mean-square error of *any* estimation procedure [15]. We would like to compute and use it to characterize the limits of achievable performance of target quantification in real-time microarrays. Assuming an unbiased estimator, the Cramer-Rao lower bound (CRLB) on the minimum mean-square error of estimating a parameter θ_i is

$$E \left(\hat{\theta}_i - \theta_i \right)^2 \geq [F^{-1}]_{ii}, \quad (9)$$

where the Fisher information matrix F is given by the negative of the expected value of the Hessian matrix of $\log p_{\mathbf{y}|\theta}(\mathbf{y})$. In other words, the entries of F are given by

$$F_{ij} = -E_{\mathbf{y}} \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p_{\mathbf{y}|\theta}(\mathbf{y}). \quad (10)$$

Since the expectation is over only \mathbf{y} , F (and hence the CRLB) is a function of θ .

Assuming jointly Gaussian distribution of the observation vector \mathbf{y} in (6), we may write

$$p_{\mathbf{y}|\theta}(\mathbf{y}) = \frac{1}{\sqrt{(2\pi)^{NL} |R|}} \exp \left[-\frac{1}{2} (\mathbf{y} - \bar{\mathbf{y}})^T R^{-1} (\mathbf{y} - \bar{\mathbf{y}}) \right],$$

where $\bar{\mathbf{y}}$ denotes the mean of \mathbf{y} . Therefore,

$$\begin{aligned}
\log p_{\mathbf{y}|\theta}(\mathbf{y}) &= -\frac{NL}{2} \log(2\pi) - \frac{1}{2} \log \det R \\
&- \frac{1}{2} (\mathbf{y} - \bar{\mathbf{y}})^T R^{-1} (\mathbf{y} - \bar{\mathbf{y}}).
\end{aligned}$$

Note that both $\bar{\mathbf{y}}$ and R are function of θ . Let us denote $L(\theta) = \log(p_{\mathbf{y}|\theta}(\mathbf{y}))$, $L_1(\theta) = \log \det R$, and $L_2(\theta) = (\mathbf{y} - \bar{\mathbf{y}})^T R^{-1} (\mathbf{y} - \bar{\mathbf{y}})$, so that we can write

$$L(\theta) = -\frac{NL}{2} \log(2\pi) - \frac{1}{2} L_1(\theta) - \frac{1}{2} L_2(\theta).$$

Therefore, the Fisher information matrix can be written as

$$F = \frac{1}{2} E_{\mathbf{y}} \{ \nabla_{\theta} \nabla_{\theta}^T L_1(\theta) \} + \frac{1}{2} E_{\mathbf{y}} \{ \nabla_{\theta} \nabla_{\theta}^T L_2(\theta) \}.$$

The components of $\nabla_{\theta} \nabla_{\theta}^T L_1(\theta)$ are given by

$$\frac{\partial^2 L_1(\theta)}{\partial \theta_i \partial \theta_j} = \text{tr} \left\{ -R^{-1} \frac{\partial R}{\partial \theta_j} R^{-1} \frac{\partial R}{\partial \theta_i} + R^{-1} \frac{\partial^2 R}{\partial \theta_j \partial \theta_i} \right\},$$

where $\text{tr} \{ \cdot \}$ denotes the trace operation over its argument. Clearly,

$$E_{\mathbf{y}} \left\{ \frac{\partial^2 L_1(\theta)}{\partial \theta_i \partial \theta_j} \right\} = \frac{\partial^2 L_1(\theta)}{\partial \theta_i \partial \theta_j}.$$

Finding components of $\nabla_{\theta} \nabla_{\theta}^T L_2(\theta)$ is more involved. First, note that

$$\begin{aligned}
\frac{\partial}{\partial \theta_i} L_2(\theta) &= - \left(\frac{\partial \bar{\mathbf{y}}}{\partial \theta_i} \right)^T R^{-1} (\mathbf{y} - \bar{\mathbf{y}}) \\
&- (\mathbf{y} - \bar{\mathbf{y}})^T R^{-1} \frac{\partial R}{\partial \theta_i} R^{-1} (\mathbf{y} - \bar{\mathbf{y}}) \\
&- (\mathbf{y} - \bar{\mathbf{y}})^T R^{-1} \frac{\partial \bar{\mathbf{y}}}{\partial \theta_i}.
\end{aligned}$$

The second derivative is given by

$$\begin{aligned}
\frac{\partial^2 L_2(\theta)}{\partial \theta_i \partial \theta_j} &= 2 \left(\frac{\partial \bar{\mathbf{y}}}{\partial \theta_i} \right)^T R^{-1} \frac{\partial R}{\partial \theta_j} R^{-1} (\mathbf{y} - \bar{\mathbf{y}}) \\
&- 2 \left(\frac{\partial^2 \bar{\mathbf{y}}}{\partial \theta_i \partial \theta_j} \right)^T R^{-1} (\mathbf{y} - \bar{\mathbf{y}}) \\
&+ 2 \left(\frac{\partial \bar{\mathbf{y}}}{\partial \theta_i} \right)^T R^{-1} \frac{\partial \bar{\mathbf{y}}}{\partial \theta_j} \\
&+ 2 (\mathbf{y} - \bar{\mathbf{y}})^T R^{-1} \frac{\partial R}{\partial \theta_i} R^{-1} \frac{\partial R}{\partial \theta_j} R^{-1} (\mathbf{y} - \bar{\mathbf{y}}) \\
&+ 2 (\mathbf{y} - \bar{\mathbf{y}})^T R^{-1} \frac{\partial R}{\partial \theta_i} R^{-1} \frac{\partial \bar{\mathbf{y}}}{\partial \theta_j} \\
&- (\mathbf{y} - \bar{\mathbf{y}})^T R^{-1} \frac{\partial^2 R}{\partial \theta_i \partial \theta_j} R^{-1} (\mathbf{y} - \bar{\mathbf{y}}).
\end{aligned}$$

Therefore, $E_{\mathbf{y}} \left\{ \frac{\partial^2 L_2(\theta)}{\partial \theta_i \partial \theta_j} \right\} =$

$$\begin{aligned}
&= 2 \left(\frac{\partial \bar{\mathbf{y}}}{\partial \theta_i} \right)^T R^{-1} \frac{\partial \bar{\mathbf{y}}}{\partial \theta_j} + 2 \text{tr} \left\{ R^{-1} \frac{\partial R}{\partial \theta_i} R^{-1} \frac{\partial R}{\partial \theta_j} \right\} \\
&- \text{tr} \left\{ R^{-1} \frac{\partial^2 R}{\partial \theta_i \partial \theta_j} \right\}
\end{aligned}$$

Combining all of the above,

$$F_{ij} = \left(\frac{\partial \bar{\mathbf{y}}}{\partial \theta_i} \right)^T R^{-1} \frac{\partial \bar{\mathbf{y}}}{\partial \theta_j} + \frac{1}{2} \text{tr} \left\{ R^{-1} \frac{\partial R}{\partial \theta_i} R^{-1} \frac{\partial R}{\partial \theta_j} \right\}.$$

Therefore, to evaluate F_{ij} , we need to compute R , $\partial R / \partial \theta_i$, and $\partial \bar{\mathbf{y}} / \partial \theta_i$. Now, computation of the covariance matrix R is outlined in the previous section. On the other hand, the (l_1, l_2) block of $\partial R / \partial n_t$ is

$$\frac{\partial R_{l_1 l_2}}{\partial n_t} = H \frac{\partial R_{\mathbf{u}, l_1 l_2}}{\partial n_t} H^T,$$

where $R_{\mathbf{u}, l_1 l_2}$ is computed from (8). Note that

$$\frac{\partial A}{\partial \theta_i} = \frac{1}{n_p} \frac{\partial D_{\mathbf{b}}}{\partial \theta_i} [I \otimes (\mathbf{1}^T)], \quad \frac{\partial D_{\mathbf{b}}}{\partial \theta_i} = D_{e_i \otimes \mathbf{k}_i},$$

where e_i denotes the N -dimensional unit vector with all zero entries except the i^{th} one. Moreover, note that

$$\frac{\partial \bar{\mathbf{y}}(l)}{\partial \theta_i} = H(e_i \otimes \mathbf{k}_i) l \Delta.$$

4. A CASE STUDY

Consider an artificial example where there are M targets, each binding to its corresponding probe with hybridization reaction rates $k_{ii} = \alpha$. Moreover, each target is binding to all $N - 1$ non-specific probes with cross-hybridization rates $k_{ij} = \beta$, $i \neq j$. Furthermore, assume that the concentrations of all M targets are identical, i.e., $n_i(0) = n_t$, for $i = 1, \dots, M$. This symmetric setup makes the required computations towards the CRLB straightforward, while it provide intuition for a more general scenario.

The CRLB for the real-time microarray system can be computed numerically following the procedure outlined in the previous section. On the other hand, the Fisher Information Matrix for a conventional DNA microarray which acquires a single data point in the steady state of the hybridization reaction was computed in [14] as

$$F = Q^T \Sigma_s^{-1} Q + \frac{1}{2} (Q - Q \odot Q)^T \Sigma_s^{-2} (Q - Q \odot Q), \quad (11)$$

where the (i, j) entry of Q , q_{ij} , denotes the probability that in the steady state target j is bound to probe i . Basically, the q_{ij} are computed from the steady-state probabilities of the Markov chains illustrated in Fig. 2. Moreover, Σ_s is a diagonal matrix with entries

$$\Sigma_s(i, i) = \sigma_v^2 + \sum_{j=1}^M q_{ij} (1 - q_{ij}) n_j(0).$$

For the special scenario considered in this section, inverse of the Fisher Information Matrix (11) is computed as

$$F^{-1} = \frac{1}{a - b} \left[I_M - \frac{\mathbf{1} \cdot b \cdot \mathbf{1}^T}{a + (M - 1)b} \right],$$

where

$$\begin{aligned} a &= \frac{p^2}{\sigma_s^2} + \frac{p^2(1-p)^2}{2\sigma_s^2} + (M-1) \left[\frac{q^2}{\sigma_s^2} + \frac{q^2(1-q)^2}{2\sigma_s^2} \right] \\ b &= \frac{2pq}{\sigma_s^2} + \frac{p(1-p)q(1-q)}{\sigma_s^2} \\ &+ (M-2) \left[\frac{q^2}{\sigma_s^2} + \frac{q^2(1-q)^2}{2\sigma_s^2} \right], \end{aligned}$$

p and q are the steady state probabilities of the hybridization and cross-hybridization, and

$$\sigma_s = (\sigma^2 + p(1-p)n_t + (M-1)q(1-q)n_t).$$

To illustrate how the CRLB for the real-time microarray system compares with the CRLB for the conventional microarray system, we consider an example with $N = 20$, $k_b = 0.1$, $k_c = 0.01$, $n_p = 10000$, $n_t = 1000$, $\sigma_v^2 = 1000$, $\Delta = 1$. Fig. 3 shows the binding process when a single target is present in the system.

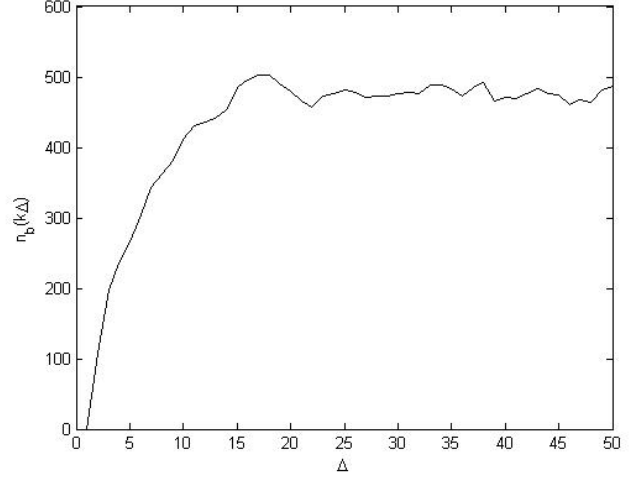


Figure 3: The binding process for a single target-probe pair.

In Fig. 4, the CRLB for the real-time microarray is compared to the CRLB for the conventional microarray when there are $M = 5$ targets present in the system. Clearly, a few temporal samples collected in the early phase of the binding process are sufficient to enable real-time system to have lower CRLB than the conventional ones.

Fig. 5 compares the CRLB of the two systems for varying number of targets, where the real-time system collects $L = 10$ samples from the early phase of the hybridization reaction. Evidently, the CRLB for the real-time system is consistently significantly lower than that for the conventional system.

5. CONCLUSION

We derived the Cramer-Rao lower bound on the mean-square error of estimating the target amounts in real-time DNA microarrays, and compared it to that of conventional microarrays. The results suggest that a few temporal samples collected in the early phase of the binding process are often sufficient to provide significant performance improvement of the real-time microarrays over the conventional ones.

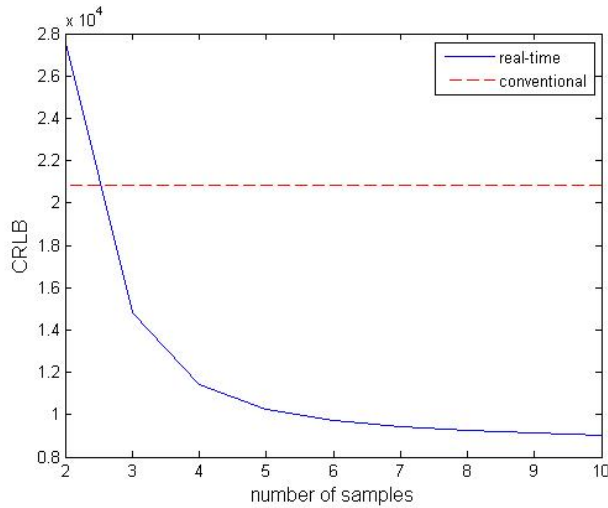


Figure 4: The CRLB for real-time vs. conventional microarray, $M = 5$ targets present in the system.

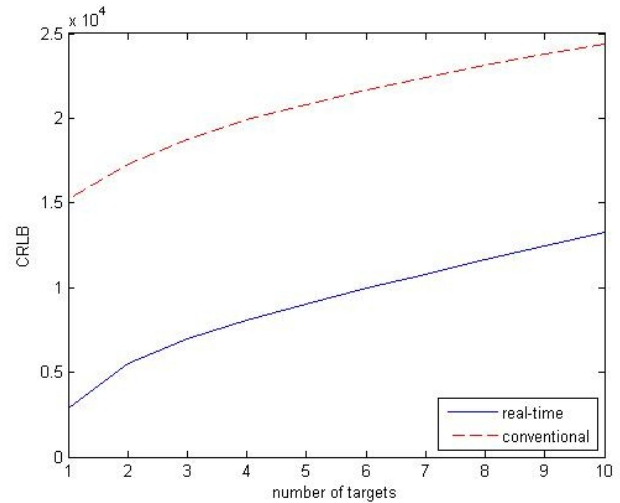


Figure 5: The CRLB for real-time vs. conventional microarray as a function of the number of targets in the system.

6. REFERENCES

- [1] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, "Quantitative monitoring of gene expression patterns with a complementary DNA microarray," *Science*, 270(5235), October 1995, pp. 467-70.
- [2] D. Shalon, S. J. Smith, and P. O. Brown, "A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization," *Gen. Research*, 6(7), July 1996, pp. 639-45.
- [3] M. Schena, *Microarray Analysis*, John Wiley & Sons, 2003.
- [4] U. R. Mueller and D.V. Nicolau (Eds.), *Microarray Technology and Its Applications*, Springer, Berlin, Germany, 2005.
- [5] E. Marshall, "Getting the noise out of gene arrays," *Science*, 306(5696):630-631, (2004).
- [6] P. K. Tan et. al., "Evaluation of gene expression measurements from commercial microarray platforms," *Nucleic Acids Res*, 31(19):5676-5684, (2003).
- [7] S. Draghici, P. Khatri, A. C. Eklund, and Z. Szallasi, "Reliability and reproducibility issues in DNA microarray measurements," *Trends Genet.*, 22(2): 101109, February 2006.
- [8] D. I. Stimpson et. al., "Real-time detection of DNA hybridization and melting on oligonucleotide arrays by using optical wave guides," *Proc. Natl. Acad. Sci. USA*, vol. 92, July 1995, 6379-6383.
- [9] M. R. Henry, P. W. Stevens, J. Sun, and D. M. Kelso, "Real-time measurements of DNA hybridization on microparticles with fluorescence resonance energy transfer," *Analyt. Bioch.*, no. 276, 1999, 204-214.
- [10] J. Bishop, A. M. Chagovetz and S. Blair, "Kinetics of multiplex hybridization: Mechanisms and implications," *Biophysical Journal*, 94:1726-1734, 2008.
- [11] V. M. Mirsky, "Affinity sensors in non-equilibrium conditions: highly selective chemosensing by means of low selective chemosensors," *Sensors*, vol. 1, 2001, pp: 13-17.
- [12] H. Vikalo, B. Hassibi, and A. Hassibi, "Modeling and estimation for real-time microarrays," *IEEE J. of Sel. Topics in Sig. Proc., Special Issue on Genomic and Proteomic Signal Processing*, vol. 2, no. 3, 2008, pp: 286-296.
- [13] A. Hassibi, H. Vikalo, J.-L. Reichmann, and B. Hassibi, "Real-time DNA microarray analysis," *Nucleic Acids Research*, doi:10.1093/nar/gkp675, 1-12, August 2009.
- [14] H. Vikalo, B. Hassibi, and A. Hassibi, "A statistical model for microarrays, optimal estimation algorithms, and limits of performance," *IEEE Transactions on Signal Processing, Special Issue on Genomics Signal Processing*, vol. 54, no. 6, June 2006, pp. 2444-2455.
- [15] H. Cramer, *Mathematical Models of Statistics*, Princeton University Press, Princeton, NJ 1946.